LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

UCRL-TR-201427

STK Bladestore Tests

Todd Heer

December 12, 2003

Disclaimer

This document was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor the University of California nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or the University of California, and shall not be used for advertising or product endorsement purposes.

This work was performed under the auspices of the U.S. Department of Energy by University of California, Lawrence Livermore National Laboratory under Contract W-7405-Eng-48.

STK Bladestore Tests

August-September 2003

Overview

The STK Bladestore is a disk subsystem consisting of ATA disks, fiber channel connectivity, and a RAID controller (LSI manufactured). There are essentially four host connections* and four backend fiber connections. The host side ports are 2Gb/sec and with their advertised 400MB/sec bandwidth, the disk side ports are 1GB/sec.

Our goal is to test this flavor of disk to see what the real world performance might be.

*Actually 8 ports across four cards, but they recommend to use only one port on each card as those are basically mini-hubs

Hardware

The borrowed STK Bladestore consisted of 2 "B150" disk drawers each with 10 blades of five 250GB drives for a total raw capacity of just under 25TB in two drawers, one "BC84"Control Module, and the "F40" 40U height cabinet.

In 16U of height in a 40U cabinet, there was 567 lbs of hardware. That is something to make note of. A full rack would weigh in at 1147 pounds.

The host system is an IBM 7026-6M1 with 8GB of RAM, 8 PowerPC_RS64 IV processors, and four separate RIO drawers. The HBAs are IBM 6228 fiber channel cards which are 2 gigabit per second (gb/s) speed capable. The disks were connected natively to the 6M1 without the use of a fiber switch (such as a Brocade). This system has sustained I/O throughputs of over 600MB/sec, so it was more than adequate for this test. It is also representative of what we run in our production storage environment.

The operating system is AIX 5.1 with maintenance level 3 applied.

Limitations

Connected natively as we were, there is an apparent two-path limitation from the host. This means that I would only be able to test two HBAs worth of throughput. Although I haven't verified it, my suspicion is that this would be alleviated in a SAN environment – perhaps a Brocade fiber switch in-between the disk subsystem and the host. IBM support verified this. Given more time, I would test that assumption.

It should also be noted that it takes quite a bit of time to reconfigure/format new RAID groups in the Bladestore, as the disk is quite dense for its size. We first configured 2TB LUNs and that took over 18 hours to format.

STK Bladestore Tests

August-September 2003

Tests

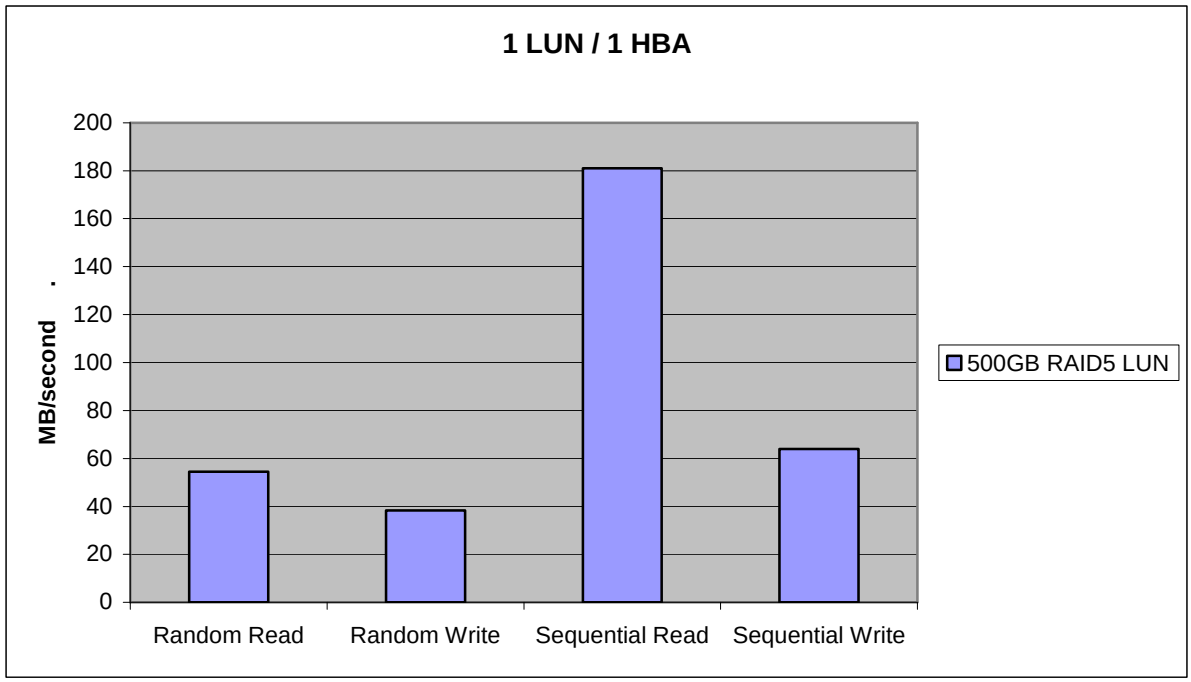
All runs have 2 types of tests run against them: random access and sequential. The random access tests take a LUN, splits it into 16 logical volumes, and then hammers away (reads or writes depending on which part of the test is being run) with 4MB buffers at any point in the logical volume for 10 minutes (5 minutes for writes then 5 minutes for reads). It is know as “donnie.”

The sequential test is simply the dd command with 16MB buffers. Writes are simply a string of zero’s (/dev/zero). They start from the beginning of the LUN and/or logical volume (numbers have proven to be the same either way) and are allowed to run for a period of time sufficient to garner good numbers.

The Hard Facts

All tests were run against 500GB LUNs. This is the LUN size that seemed to perform the best, was able to be recognized by AIX (2TB LUNs weren’t able to be configured), and offer enough size to be potentially useful without being too large for our intended purpose. The segment size on the controller was observed to be 128KB.

The first run was a single LUN over a single path. Only one hdisk is accessed, making use of one HBA. As there is no contention between disks or HBAs, this test’s results are quite useful as a baseline and aid in determining where bottlenecks lay in future tests.

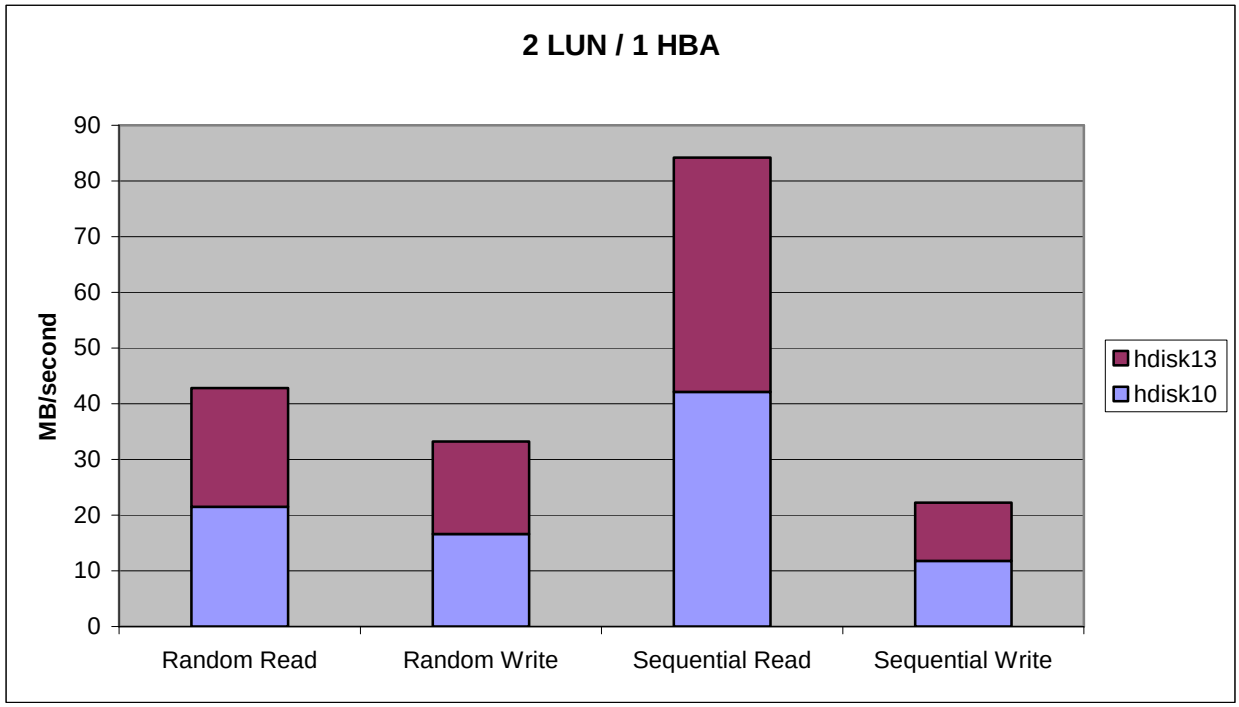


Run 1

In this test we see that the random reads are around 54MB/sec and random writes are 38MB/sec. Sequential reads are impressive at 181MB/sec, which is near the theoretical limit of 2GB/sec line speeds. Sequential writes are low at 64MB/sec. It was observed that

the first few seconds (<= 4) of a write test were approximately 50% faster than any other time in the remainder of that specific run. I suspect this is a result of caching and is really not a true measure of sustained write performance.

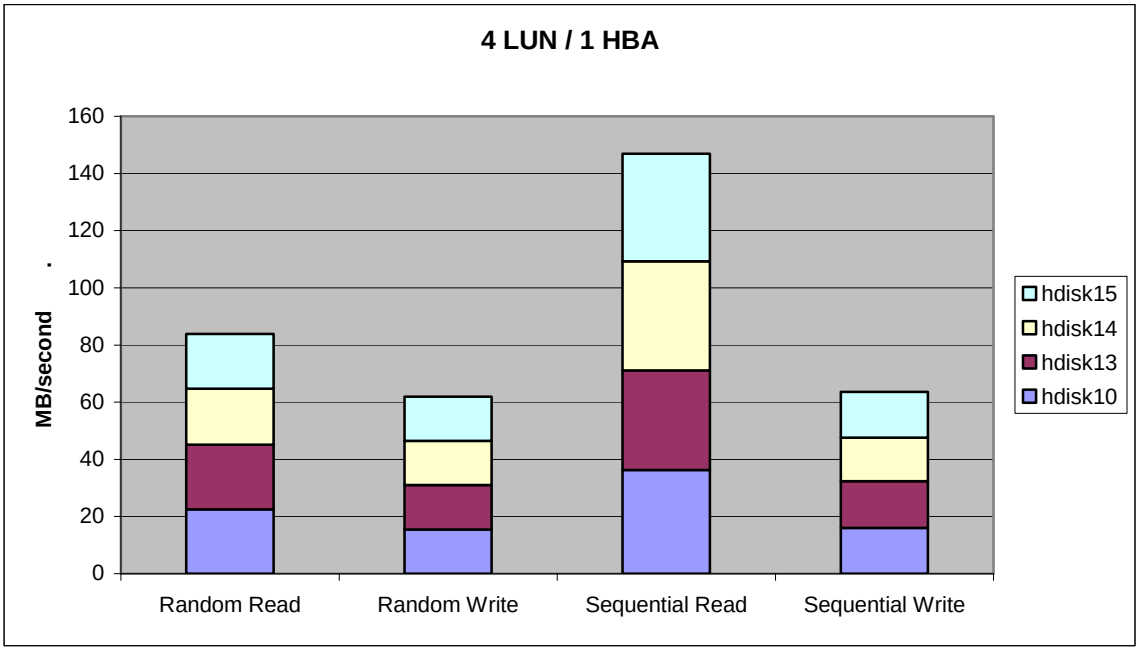
The second run is two LUNs over one HBA. Ideally we would expect to see a twofold increase in throughput for each test.



Run 2

We see that not only random reads drop per LUN, but the aggregate is even lower. This is also true for random writes, sequential reads, and sequential writes.

The third run is four LUNs over one HBA. Ideally, we would observe performance numbers be either a multiple of the number of LUNs over the baseline tests or be near the theoretical limit of 2Gb/sec fiber speeds (which is somewhere around 200MB/sec), whichever is lower.

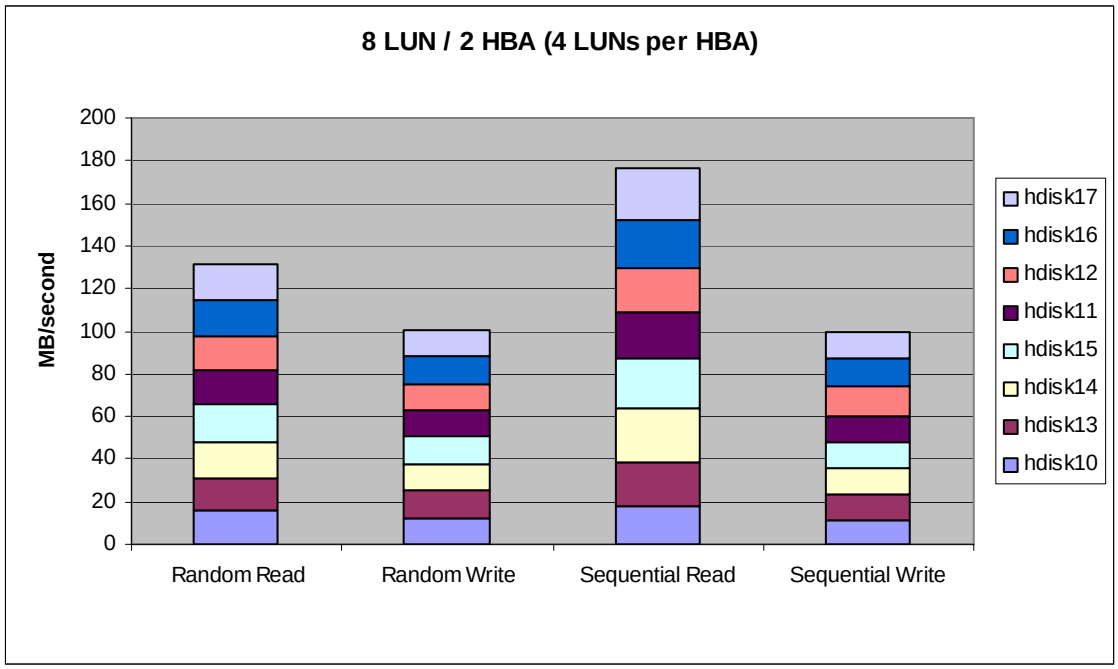


Run 3

Random reads were observed to be an aggregated 83.8 MB/sec. This is about a 62% drop from what we might expect to see if it were a linear improvement with 4 LUNs (4 * 54.4 or 2Gb/sec speeds). Random writes came in at nearly 62MB/sec. This also shows about the same 60% drop in what might be expected linearly. At this point we can see the caveat about serial ATA disk in the numbers: they are not positioned to be high I/O per second capable disk subsystems. With roughly four times the number of I/O requests coming into the controller, we see 60% loss of single LUN performance (but still achieving a higher aggregated number – unlike Run 2).

Sequential reads fell off a bit to 147MB/sec from the single test of 181MB/sec (a 20% drop in overall performance). Again I suspect the juggling of four request streams has something to do with this. Sequential writes appear to be unaffected at 63.6MB/sec – nearly the same as a single sequential write. It appears the controller doesn't have a problem with I/O's when it's only being fed at 64MB/sec.

The fourth run is 8 LUNs over 2 HBAs. This test is useful because we can determine where a bottleneck is between the controller and host port cards on the controller (fiber runs).



Run 4

Random reads performed at an aggregate rate of 131MB/sec. This is the culmination of two HBAs performing at 65.6 and 65.4 MB/sec respectively. We see again the reduction of expected performance of the previous test – a 21% drop over the 166MB/sec we'd expect to see by doubling our multiple LUN / single HBA test. However, the aggregate 131MB/sec is the most we've seen for the random read test.

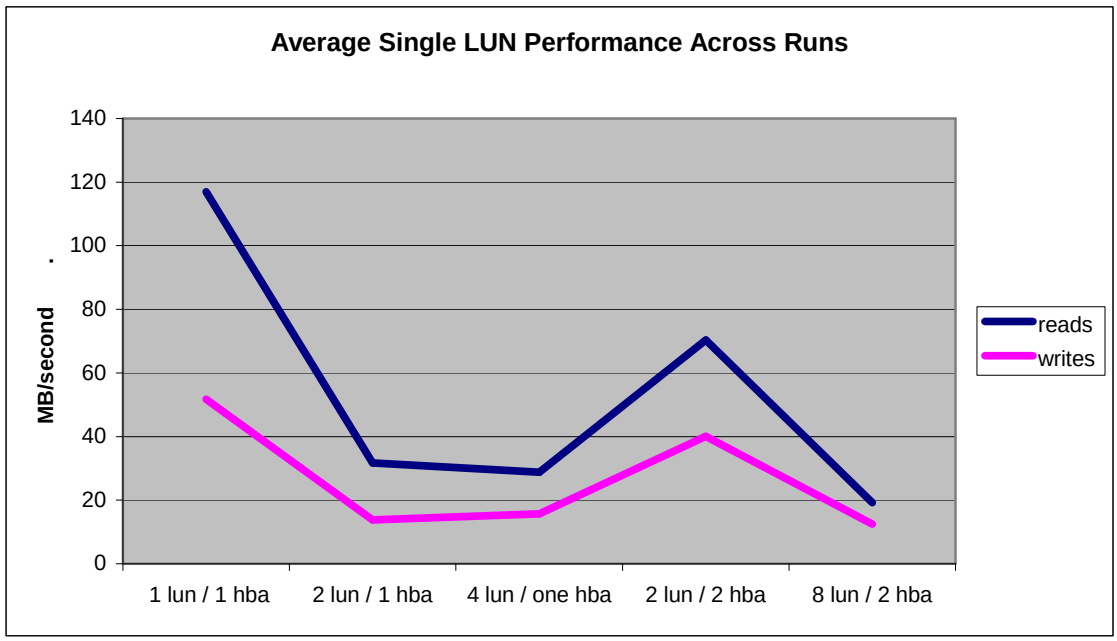
Random writes performed at 100 MB/sec. This is the culmination of two HBAs performing at 50.3 MB/sec each. This also shows a similar 18% drop in expected performance we'd like to see over the multiple LUN / single HBA test of 124MB/sec (62 * 2).

Sequential Reads came in 176.75 MB/sec. This is better than the sequential reads of the multiple LUN / single HBA test, but it's nearly the same (aggregated) as the single LUN / single HBA test. This would seem to suggest that over two host ports, the limit is somewhere around 180 MB/sec on reads. This is in line with manufacturer claims of 400MB/sec, if you assume that the two unused ports would also be able to run at that speed. However, if one host port can perform at 146MB/sec on multiple LUN sequential reads, it seems logical that two host ports should have been able to run at 292 MB/sec (2x) since it's still under the 400MB/sec claimed throughput. So we must not assume the ability of 400MB/sec for any kind of test from this controller / disk configuration combo. We were limited by the use of 2 drawers of disk, rather than the 4 we initially hoped for. With 4 drawers, we would have the use of more backend channels.

Sequential Writes were 99.75 MB/sec. Again we see the 21% drop in expected performance (by doubling the previous test's 63.6 MB/sec).

Summation

Let's take a look at average LUN performance for each of the tests performed.



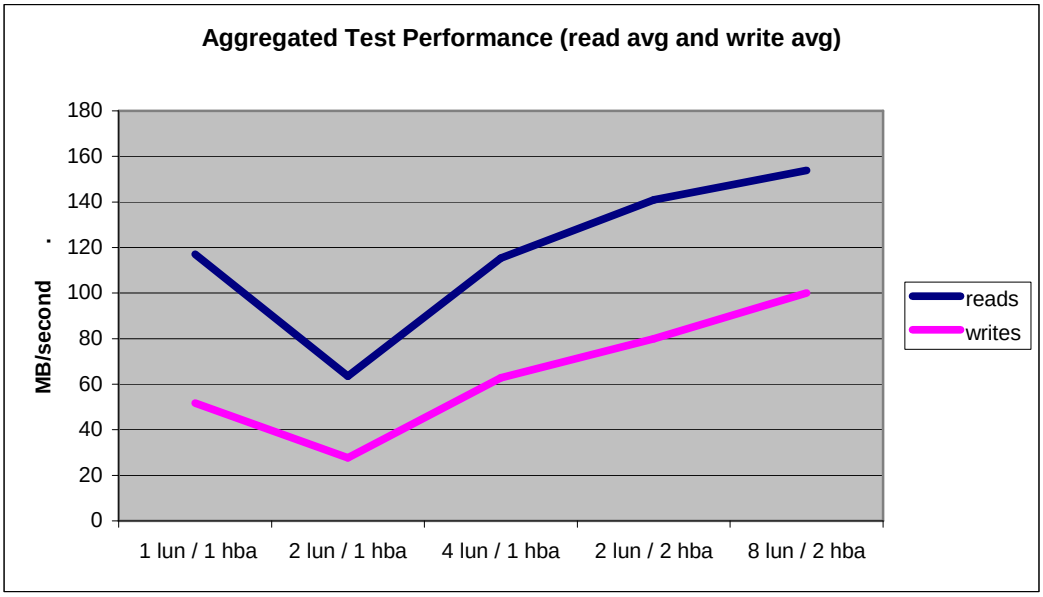
An additional test was run for this chart – 2 LUNs over 2 HBAs. You would expect to see a 2x improvement over 1 LUN/1 HBA, again this proves to be not the case.

Reads and writes were averaged together for simplicity. That is, the sequential read values and random read values were averaged together to come up with one number for each test. It's apparent, as one studies this graph, that any kind of multiple requests on either read or write causes performance to drop. For use in HPSS, we might not want more than one request stream per controller.

And finally, here is a chart for the aggregated performance for each run.

STK Bladestore Tests

August-September 2003



Again reads were averaged and writes were averaged. This chart shows an upward trend as you add fiber channel cards. I'd like to see a four drawer Bladestore configuration utilizing all four backend and four front-end ports. Beyond what is represented here is speculation.

Other less tangible things such as resistance to corruption, ease of maintenance, availability, and others were not tested due to time constraints. Testing was limited to performance runs in order to determine real-world MB/second numbers.

The point to take away from this suite of tests is that for an application that does not need more than one stream and performs sequential I/O (with emphasis on reading over writing), this disk subsystem may represent significant bang-for-the-buck. However in multiple stream scenarios where high performance I/O numbers are expected (line speed of 2Gb/sec, for example), this system falls way short of the mark.

Todd Heer
theer@llnl.gov
09-15-2003